

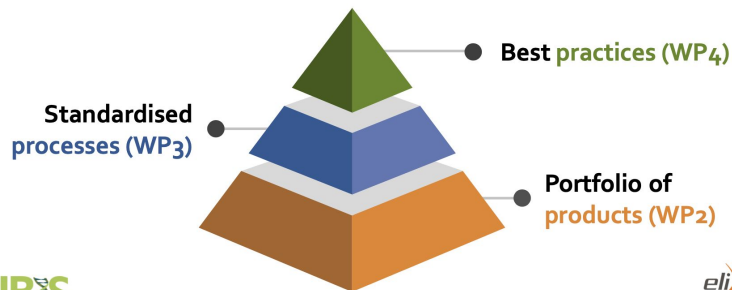
The future of data integration? Shared recipes and templates for FAIR data, tools, and services

Wolmar Nyberg Åkerström
NBIS Data Management Unit

The future of data integration?

ELIXIR Interoperability Platform

Connecting processes to products/practices



NBS



FAIRify your data for yourself

This workshop will showcase how to put the FAIR principles into practice and is specifically designed to cater to life science researchers at all career stages. The FAIR principles are intended to support effective data sharing up to the degree that you can assert "it just works with my software". But they apply equally well to the social aspects of data sharing and can serve as a guide to help you address the question "if I leave the lab today, how can my colleagues understand, find and use my data?".

Learning outcomes:

- Understanding of why funders promote the FAIR principles and how FAIR is useful to you in projects and collaborations of any size
- Experience of using a life cycle perspective to identify where and when to apply the FAIR principles most effectively
- Know examples of simple strategies and effective practices to implementing the FAIR principles



ALL HANDS 2024

Go to
www.menti.com

Enter the code

8617 4959

Or use QR code



The future of data integration

Shaping ELIXIR's
Interoperability Platform
with *real-world* use cases

ELIXIR Interoperability Platform
(24-TECH-Interop)



NBS

NATIONAL BIODATA REPOSITORY

SciLifeLab

<https://nbs.se>

Publish code, software and workflows

Examples of why and how from SciLifeLab

NBIS Data Management Team
data-management@scilifelab.se

Presented by Wolmar Nyberg Åkerström
SciLifeLab Data Management seminar series Spring 2024, Online
27 March 2024



<https://doi.org/10.17044/scilifelab.25483537>



Why implement the FAIR principles?

More value to publicly funded research

Improve peer-review process

High research integrity

Better research output

Background: 'FAIR Principles'
by Martínez-Lavanchy, et al (2019),
CC-BY 4.0. doi:10.11581/dtu:00000049



Why implement the FAIR principles?

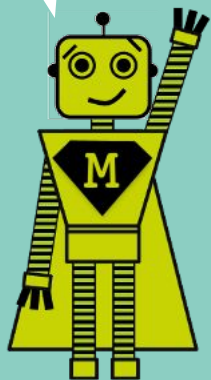
More value to publicly funded research

Improve peer-review process

Works with my software

High research integrity

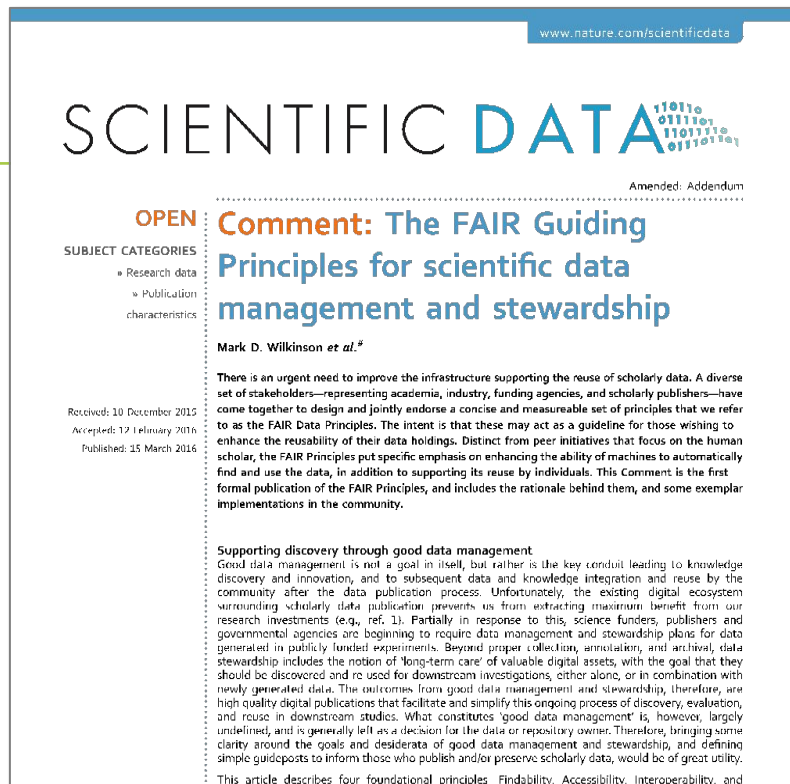
Better research output



Background: 'FAIR Principles'
by Martínez-Lavanchy, et al (2019),
CC-BY 4.0. doi:10.11581/dtu:00000049

The FAIR principles

- Promote efficient discovery and reuse by providing guidelines to make digital resources
 - ❑ Findable
 - ❑ Accessible
 - ❑ Interoperable
 - ❑ Reusable
- Address aspects enabling software and infrastructure to automatically find and integrate them



Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. doi:10.1038/sdata.2016.18

Making sense of your group's data



“ data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings”



How about another group's data?



Making sense of another group's data



“ data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings”

“ data should be assessable so that judgments can be made about their reliability and the competence of those who created them”.



How about FAIR research data?

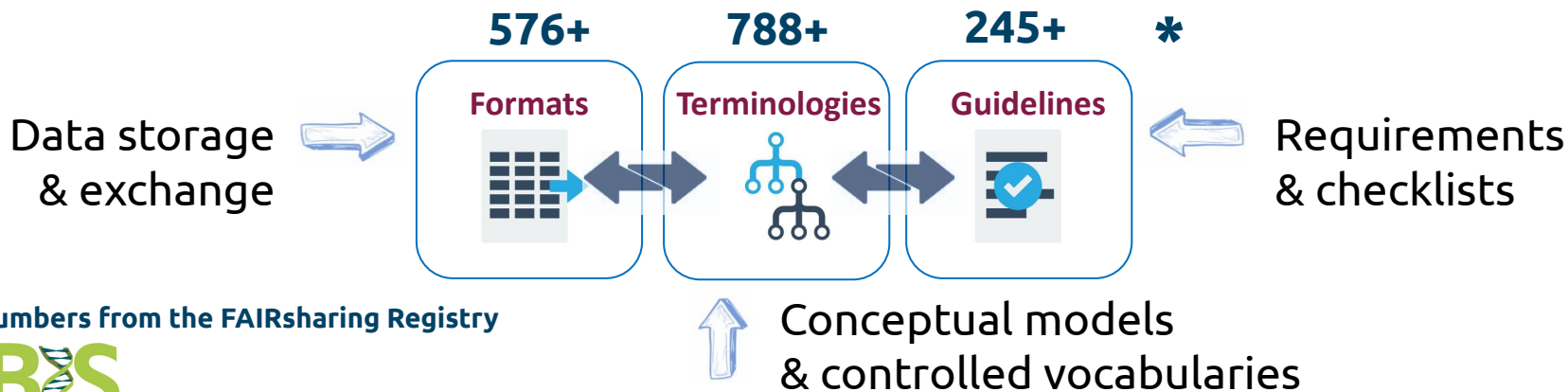


Making sense of FAIR research data



“ data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings”

“ data should be assessable so that judgments can be made about their reliability and the competence of those who created them”.



Publish code, software and workflows

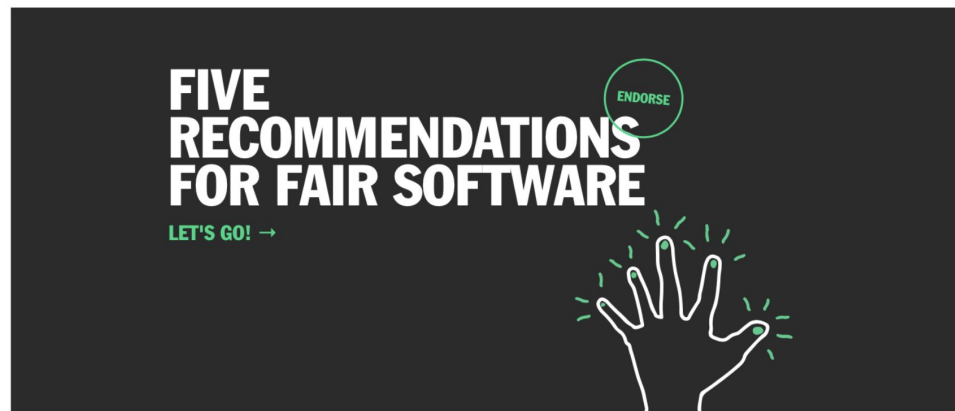
Examples of why and how from SciLifeLab

NBIS Data Management Team
data-management@scilifelab.se

Presented by Wolmar Nyberg Åkerström
SciLifeLab Data Management seminar series Spring
27 March 2024



Checklist for sharing software



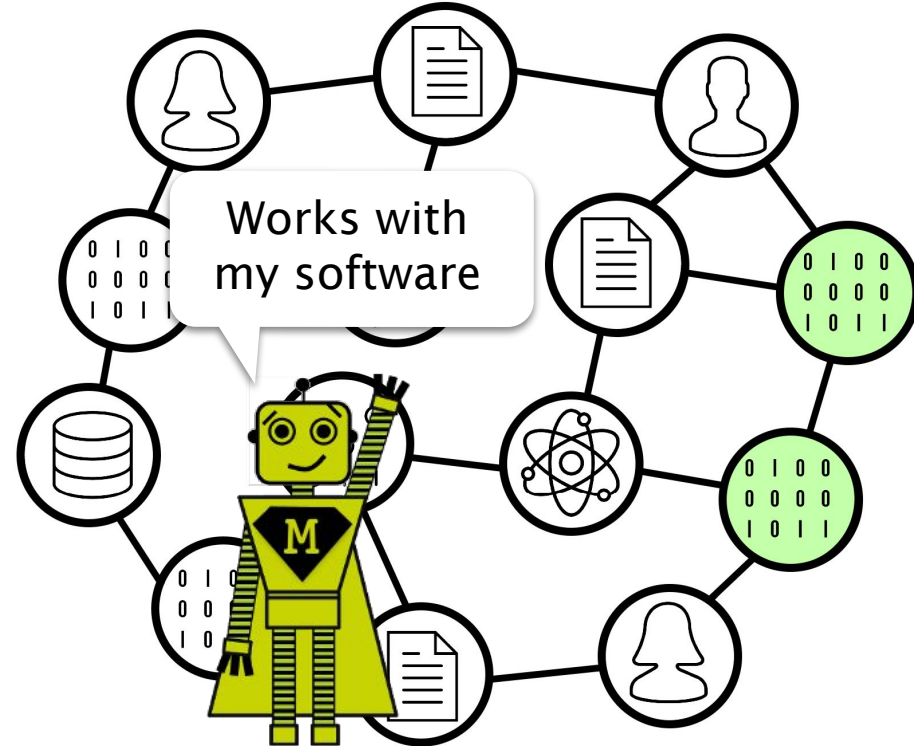
Towards a web of FAIR data and services



“From Gutenberg to Berners-Lee”

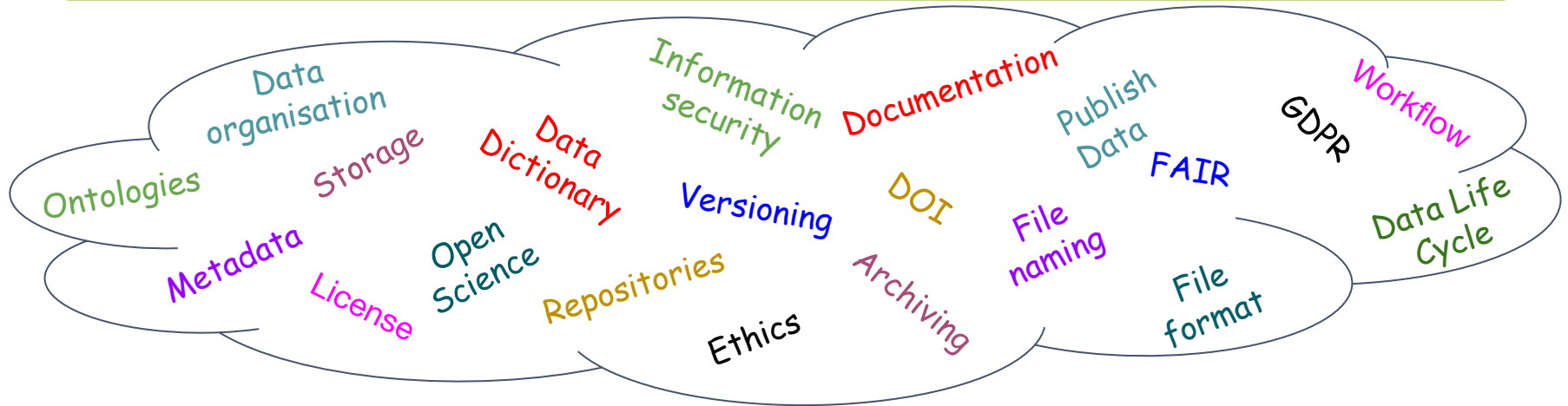
- What types of research assets?
- Where will people be looking?
- Who should have access?
- Which standards/specifications?
- What cross-references and what documentation can you provide?
(Rich metadata!)

Graph: “PID Graph” from the FREYA project



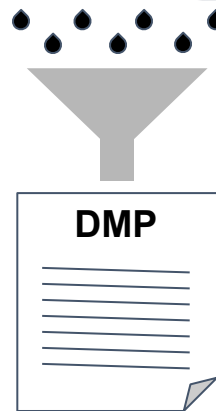
Robot: MetaManMachine by Nikola Vasiljevic (2021),
CC BY-SA 4.0, doi:10.5281/zenodo.4471098

Documenting your approach to FAIR practices



Data Management Plan (DMP):

A document addressing requirements and practices for the project's data



The Swedish Research Council: All who are awarded a grant from the Swedish Research Council must have a data management plan if the research generates research data.



Research outputs and metadata in context (I)



Study & data
design

Sampling
& specimen
collection

Sample
preparation

Sample analysis
& data generation

Data processing
to prepare inputs
for analysis

Data
analysis

Communicating
results

Procedures

data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, ...

Biosamples and instruments

populations (statistical) and inclusion criteria,
physical processing steps,
working storage conditions,
long-term storage location,
sample quality assessment,
sample annotations,
reagents, instruments, kits, ...

Data and computational workflows

digital processing steps,
working storage conditions,
long-term storage location,
data quality assessment,
sample/data annotations,
reference data,
analysis method...

Outputs

publications,
data,
tools,
workflows,
reports,
dashboards, ...

“Protocol” & “project plan” icons by Justin Blake, and “infrastructure” icon by Eko Purnomo, from thenounproject.com



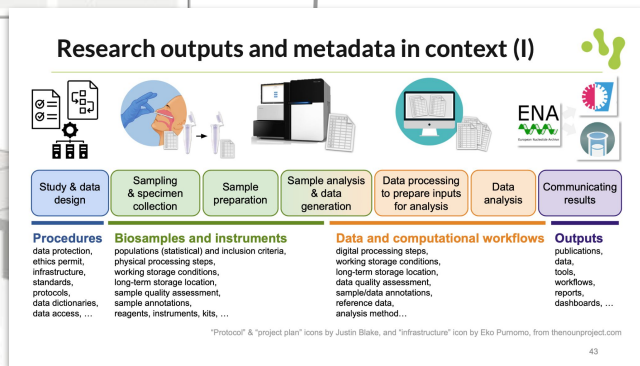
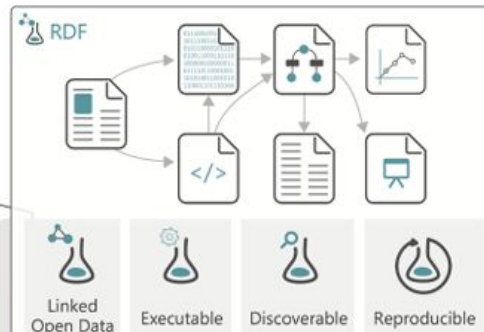
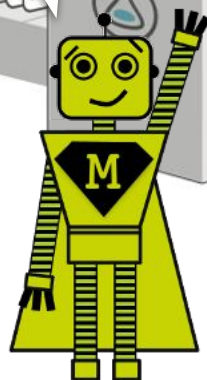
Research outputs and metadata in context (II)



 Enabling **reproducible**, transparent research.



Works with
my software



Robot: MetaManMachine by Nikola Vasiljevic (2021),
CC BY-SA 4.0, doi:10.5281/zenodo.4471098





FAIRify your data for yourself

This workshop will showcase how to put the FAIR principles into practice and is specifically designed to cater to life science researchers at all career stages. The FAIR principles are intended to support effective data sharing up to the degree that you can assert “it just works with my software”. But they apply equally well to the social aspects of data sharing and can serve as a guide to help you address the question “if I leave the lab today, how can my colleagues understand, find and use my data?”.

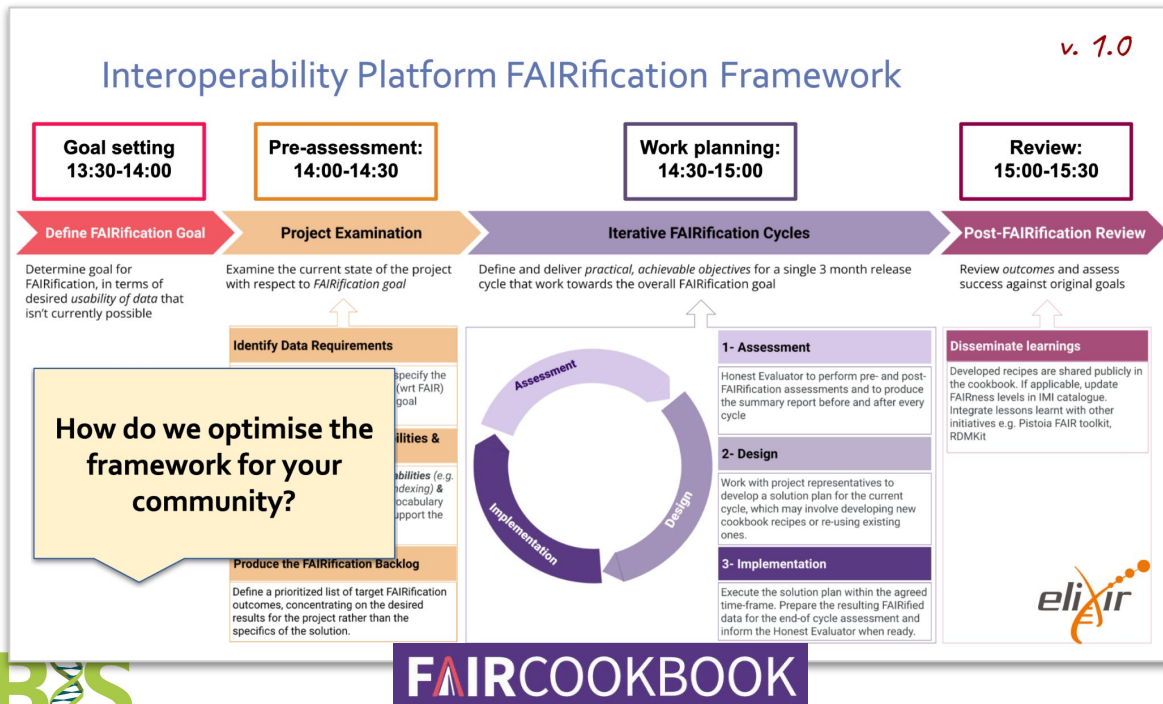
Learning outcomes:



- Understanding of why funders promote the FAIR principles and how FAIR is useful to you in projects and collaborations of any size
- Experience of using a life cycle perspective to identify where and when to apply the FAIR principles most effectively
- Know examples of simple strategies and effective practices to implementing the FAIR principles



FAIRification framework

First draft and future directions





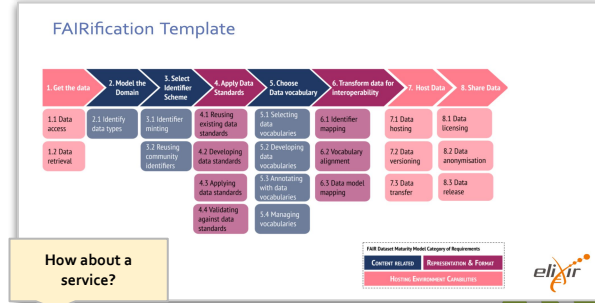
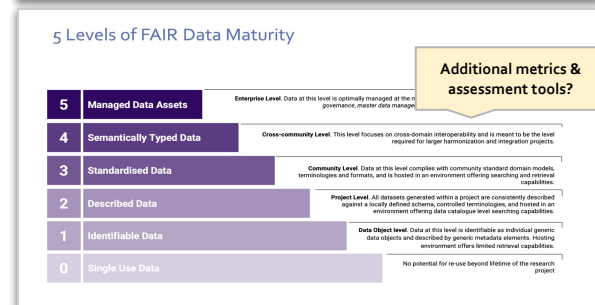
wna.se/eccb2024

Practical and pragmatic FAIRification targeting ELIXIR Communities

Tony Burdett, Ibrahim Emam, Yojana Gadiya, Nils Hoffmann, Nick Juty & Wolmar Nyberg Åkerström

 @ELIXIREurope
 company/elixir-europe

www.elixir-europe.org
 ECCB 2024, Turku, Finland

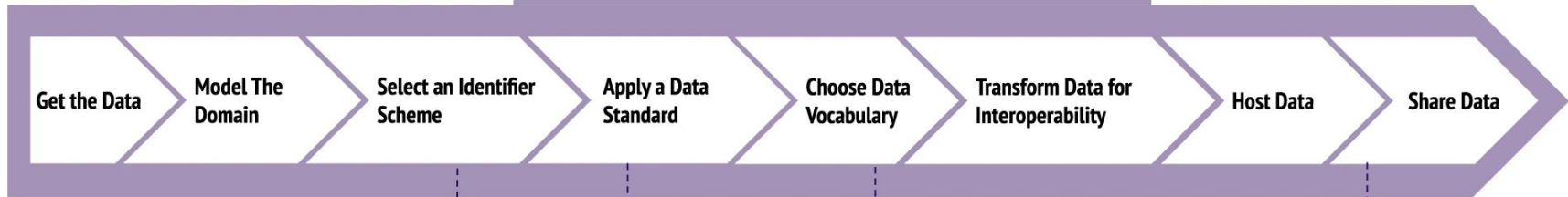


FAIR in Action Framework

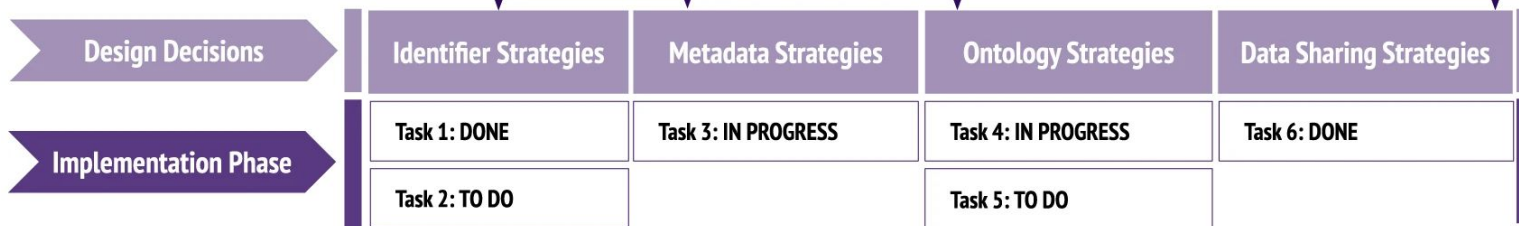
1. FAIRification Process



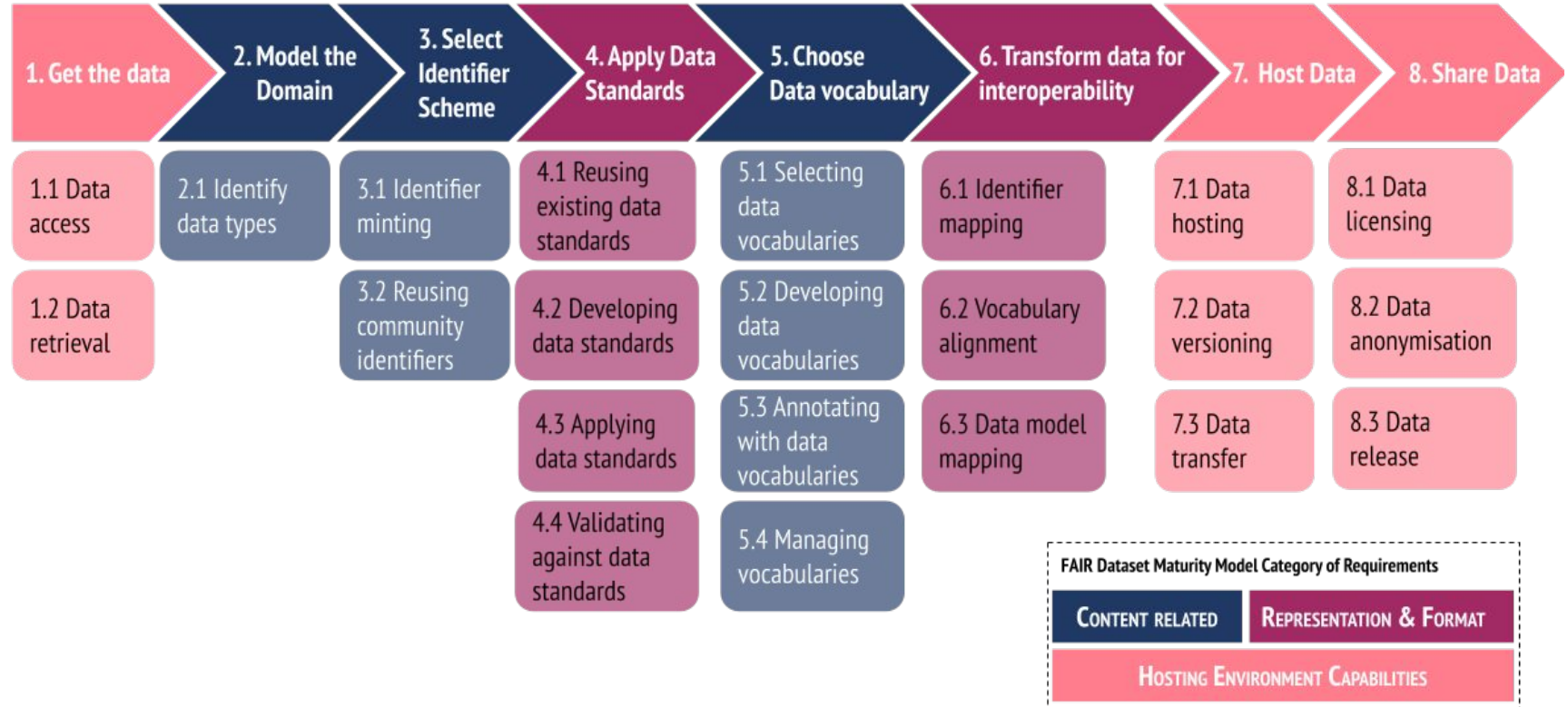
2. FAIRification Template



3. FAIRification Workplan



FAIRification Template



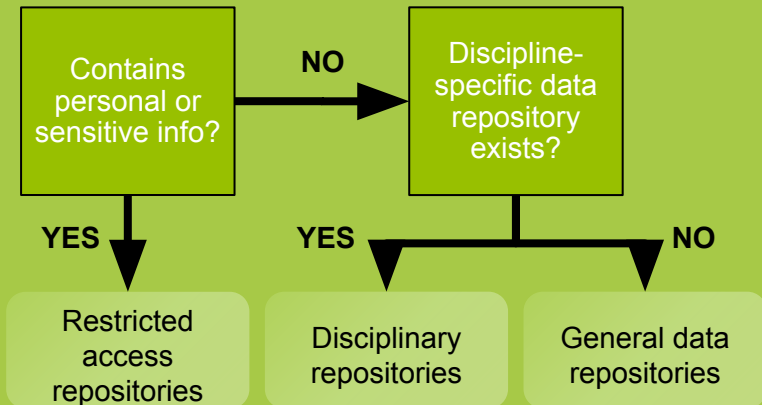


Data submission to public repositories

Making the data underlying your research publicly available to others is a fundamental part of a FAIR research process. When publicly available and appropriately described, data can be re-used by yourself as well as others. Domain-specific public repositories offer the most direct routes to making your data FAIR. This workshop will give you the why, where and how of data sharing via repository submission, including hands-on exercise. No prior knowledge is required in order to attend this workshop.

Learning outcomes:

- Know the benefits of data sharing
- Know how to find a suitable repository for different types of data
- Have experience of a repository submission





The screenshot shows a web browser window with the URL `rdmkit.elixir-europe.org/human_pathogen_genomics`. The page header includes the RDMkit logo, navigation links for 'Data management', 'About', 'Contribute', and 'GitHub', and a search bar. The left sidebar lists various data domains, with 'Human pathogen genomics' highlighted. The main content area displays the domain name 'Human pathogen genomics' with edit and delete icons, followed by an 'Introduction' section. The introduction discusses the focus on studying genetic codes of pathogens and the challenges of data management. Below this is a section titled 'Planning a study with pathogen genomic data' with a 'Description' subsection.

RDMkit

Your domain: Human pathogen genomics | RDMkit

Data management About Contribute GitHub Search RDMkit

Data management

- Data life cycle
- Your role
- Your domain
 - Bioimaging data
 - Biomolecular simulation data
 - Epitranscriptome data
 - Health data
 - Human data
 - Human pathogen genomics**
 - Intrinsically disordered proteins
 - Machine learning
 - Marine metagenomics
 - Microbial biotechnology
 - Plant sciences
 - Proteomics
 - Rare disease data

Your domain

Human pathogen genomics

On this page

Introduction

The human pathogen genomics domain focuses on studying the genetic code of organisms that cause disease in humans. Studies to identify and understand pathogens are conducted across different types of organisations ranging from research institutes to regional public health authorities. The aims can include urgent outbreak response, prevention measures, and developing remedies such as treatments and vaccines.

Data management challenges in this domain include the potential urgency of data sharing and secondary use of data across initiatives emerging from research, public health and policymakers. While pathogenic organisms are the object of interest, there are many considerations to take into account when dealing with samples collected from patients, pathogen surveillance, and human research subjects.

The genomic data can represent anything from the genetic sequence of a single pathogen isolate to various fragments of genetic materials from a flora of pathogens in a larger population. Other data can represent a wide range of contextual information about the human host, the disease, and various environmental factors.

Planning a study with pathogen genomic data

Description

While the objects of interest in this domain are pathogens, the data is usually derived from samples originating from human research subjects. This means that you must plan to either remove or handle [human data](#) during your study.

Submission to public repositories

Making your research publicly available to others is a key part of the R research process. When publicly available and shared data can be re-used by yourself as well as others, public repositories offer the most direct routes to data reuse. This workshop will give you the why, where and how of a repository submission, including hands-on practice. A small fee is required in order to attend this

of data sharing

suitable repository for different types of data
a repository submission





rdmkit.elixir-europe.org/human_pathogen_genomics

rdmkit.elixir-europe.org/data_brokering

Your tasks: Data brokering | RDMkit

RDMkit

Data management About Contribute GitHub Search RDMkit

Data management

- Data life cycle
- Your role
- Your domain
 - Bioimaging data
 - Biomolecular simulation data
 - Epitranscriptome data
 - Health data
 - Human data
 - Human pathogen genomics
 - Intrinsically disordered proteins
 - Machine learning
 - Marine metagenomics
 - Microbial biotechnology
 - Plant sciences
 - Proteomics
 - Rare disease data

Data management

- Data life cycle
- Your role
- Your domain
- Your tasks
 - Compliance monitoring
 - Costs of data management
 - Data analysis
 - Data brokering**
 - Data discoverability
 - Data management coordination
 - Data management plan
 - Data organisation
 - Data security
 - Data sensitivity
 - Data provenance
 - Data publication

Your tasks

Data brokering

On this page

Taking on the data broker role

Description

Sometimes it is challenging to exchange data across data producers, infrastructures and data sharing platforms. Some reasons can be that the data has to be pre-processed or enriched to comply with legal or organisational practices, that the data has to be translated to different data formats, or that transferring data requires expertise and access to special interfaces. By acting as a broker, you can fill this gap by negotiating a contract with data providers and/or recipients and doing the work required to make it convenient for them to exchange data.

```
graph TD
    subgraph "Individual data producer"
        A1[Prepare data/metadata] --> B1[QC/curate data]
        B1 --> C1[Analyse data]
        C1 --> D1[Store data]
        D1 --> E1[Share data]
    end
    subgraph "Data producers providing data to a data broker"
        A2[Prepare data/metadata] --> F[Transfer data (securely) to broker]
    end
    subgraph "Data recipient acting as a broker"
        F --> B2[QC/curate data]
        B2 --> C2[Analyse data]
        C2 --> D2[Store data]
        D2 --> E2[Broker data]
    end
```

Open to public

series

Openly available to others is a
process. When publicly available and
used by yourself as well as
others offer the most direct routes
will give you the why, where
commission, including hands-on
in order to attend this

Library for different types of data
mission



RDMkit

Data management

- Data life cycle
- Your role
- Your domain
 - Bioimaging data
 - Biomolecular simulation data
 - Epitranscriptome data
 - Health data
 - Human data
 - Human pathogen genomics**
 - Intrinsically disordered proteins
 - Machine learning
 - Marine metagenomics
 - Microbial biotechnology
 - Plant sciences
 - Proteomics
 - Rare disease data

Display a menu

RDMkit

Data management

- Data life cycle
- Your role
- Your domain
- Your tasks
 - Compliance monitoring
 - Costs of data management
 - Data analysis
 - Data brokering**
 - Data discoverability
 - Data management coordination
 - Data management evaluation
 - Data organisation
 - Data security
 - Data sensitivity
 - Data provenance
 - Data publication

Display a menu

rdmkit.elixir-europe.org/human_pathogen_genomics

data-guidelines.scilifelab.se/topics/data-transfer/

Data transfer | SciLifeLab Research Data Management Guidelines

SciLifeLab RDM Guidelines

Knowledge hub for the management of life science research data in Sweden

Get support About Contact

Home Research data life cycle Topics Resources

Topics

Home / Topics / Data transfer

Data transfer

Quite often large amounts of data is generated, and it can be worth spending some time considering how to transfer data from the data producer to storage and analysis environment. Consider the capacity of the internet connection, transfer via a low bandwidth network can be so time-consuming that it might be faster and easier to send the data on a hard drive through carrier services.

SciLifeLab Data Delivery System

The Data Delivery System (DDS) is a cloud-based system for the delivery of data from SciLifeLab platforms to their users. It consists of a command line interface (CLI) and a web interface. This system is e.g. used by the National Genomics Infrastructure (NGI) for delivery of sequencing data.

- [DDS homepage](#)
- [DDS documentation](#)
- [NGI guide on DDS](#)

Uppmax

Please find below some useful links from the compute resource Uppmax regarding data transfer:

- [File transfer to/from Rackham](#)
- [File transfer to/from Bianca](#)
- [NAISS-SENS Bianca Deliver user guide for NGI data](#)

Table of contents:

- [SciLifeLab Data Delivery System](#)
- [Uppmax](#)
 - [Using Aspera on Uppmax](#)
- [Transferring files over a wide range of protocols using RClone](#)
- [Resources](#)



ic

to others is a
available and
well as
direct routes
why, where
g hands-on
this

types of data

RDMkit

RDMkit

Data management

Data life cycle

Your role

Your domain

Bioimaging data

Biomolecular simulation data

Epitranscriptome data

Health data

Human data

Human pathogen genomics

Intrinsically disordered proteins

Machine learning

Marine metagenomics

Microbial biotechnology

Plant sciences

Proteomics

Rare disease data

Display a menu

RDMkit

Data management

Data life cycle

Your role

Your domain

Your tasks

Compliance monitoring

Costs of data management

Data analysis

Data brokering

Data discoverability

Data management coordination

Data management

Data organisation

Data security

Data sensitivity

Data provenance

Data publication

Display a menu



Knowledge

Topics

Home / Topics

Data

Quite often how to transfer capacity of that it might

SciLifeL

The Data Driven platforms to system is e.

- DDS he
- DDS de
- NGI gu

Uppma

Please find

- File tra
- File tra
- NAISS

Display a menu

rdmkit.elixir-europe.org/human_pathogen_genomics

main

data-submission-documentation / BioStudies / 6159-SEEC_SARS-CoV-2_variant_analysis_wastewater /

Top

6159 - Sharing SARS-CoV-2 variant analysis from Swedish wastewater samples

Submission task description

During 2021 and 2022, SLU and KTH have done monitoring of SARS-CoV-2 levels and variants in wastewater from six Swedish cities. The sequences have been published in ENA, but the variant analysis and sequencing run reports need to be submitted to BioStudies.

Procedure overview and links to examples

Links

- [Samples sheet](#) - overview of all sequenced samples and submitted data
- [file with PageTab info](#) - might be outdated
- [BioStudies - GitHub](#)
- [BioStudies-PageTab-Example](#)
- [BioStudies-PageTab-Specification](#)
- [Submit help for Biolmage Archive](#) - uses BioStudies
- [File List Guide](#) - Biolmage Archive but shows how to organise files for BioStudies in PageTab submission, where everything is written in a tsv file which is then uploaded to BioStudies
- [BioStudies database: aggregating all outputs of a life sciences study](#) - ENA training material
- [A guide to organising data associated to a publication using BioStudies](#) - ENA training material

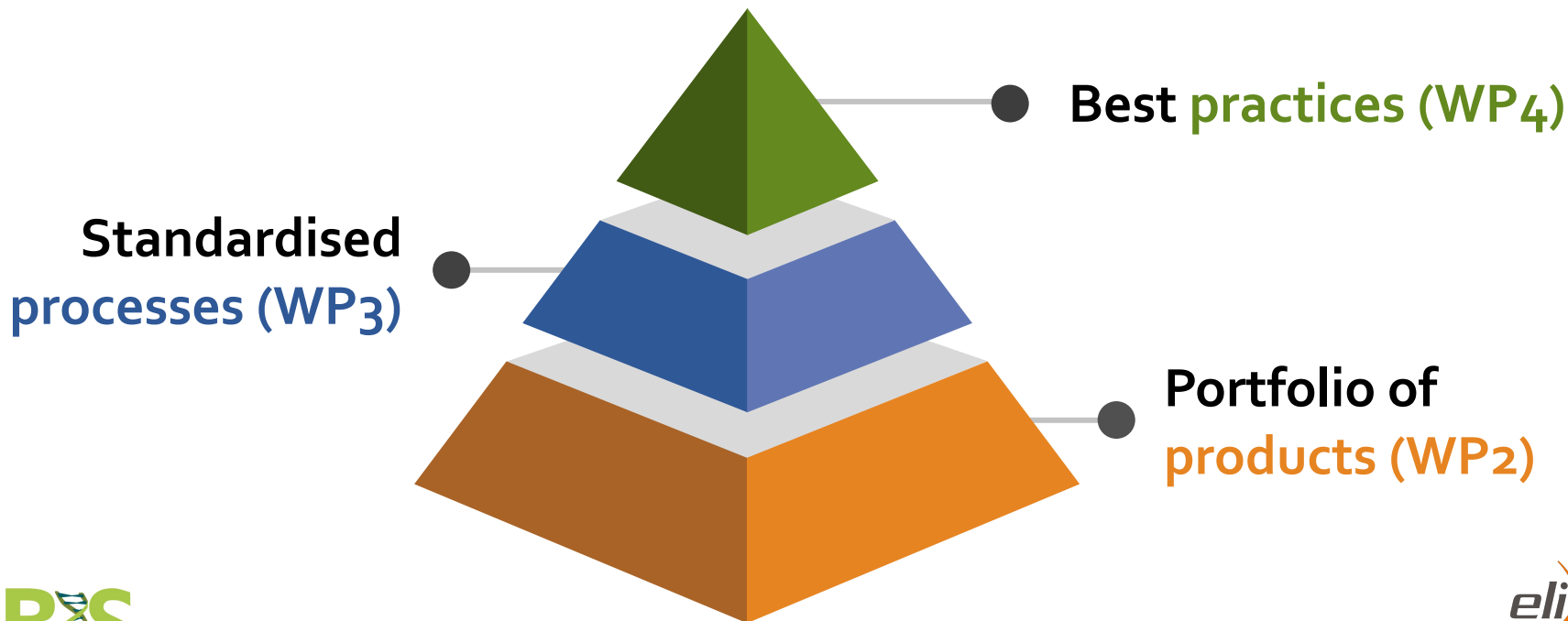
Procedure

Everything was submitted via the web browser (using PI's account at BioStudies), but the data was also submitted via the PageTab submission, where everything is written in a tsv file which is then uploaded to BioStudies (see above).

Lessons learned

ELIXIR Interoperability Platform

Connecting processes to products/practices



ELIXIR Interoperability Platform

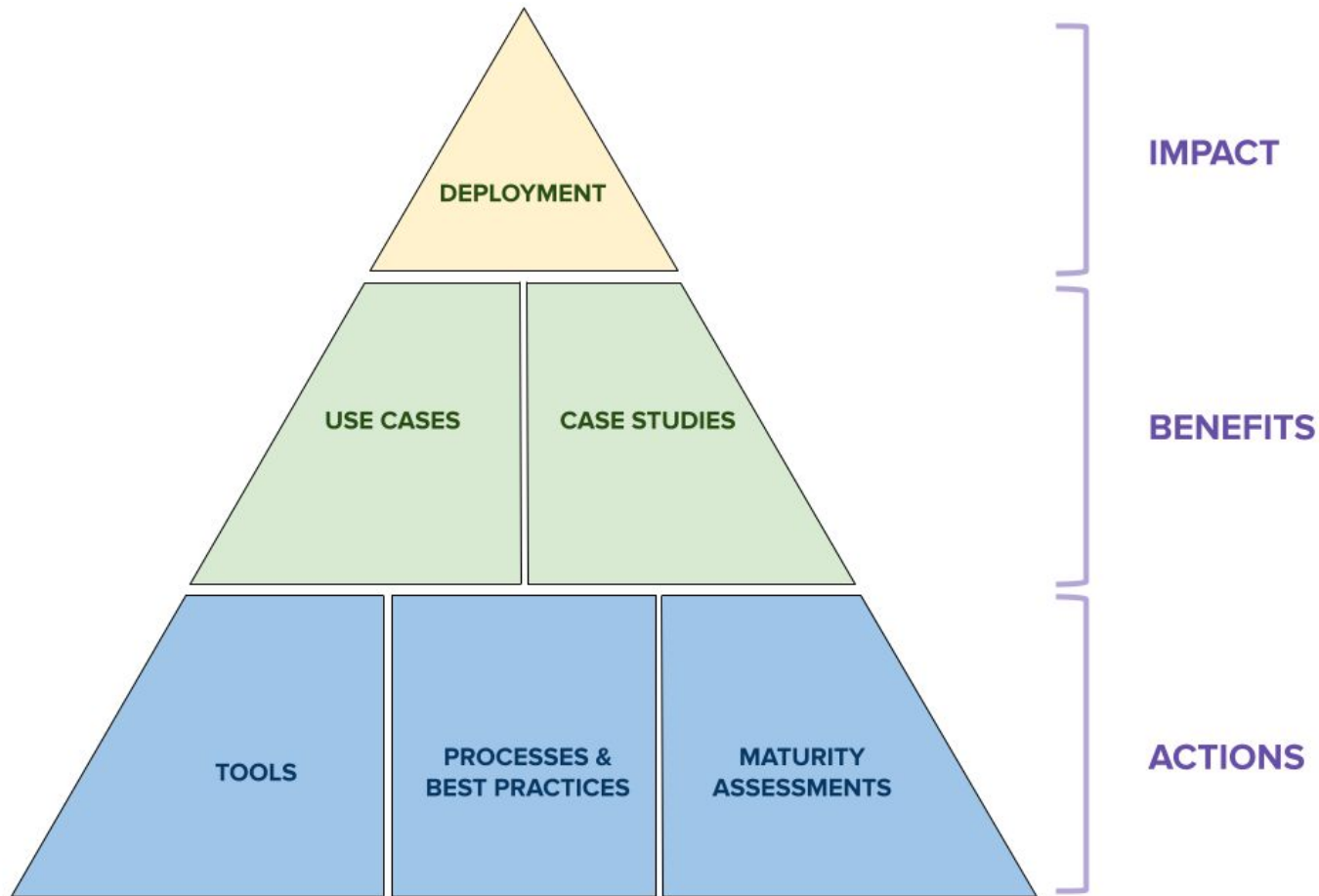
Connecting processes to products/practices

Assemble **products** into
reusable **processes** that
drive good **practice**

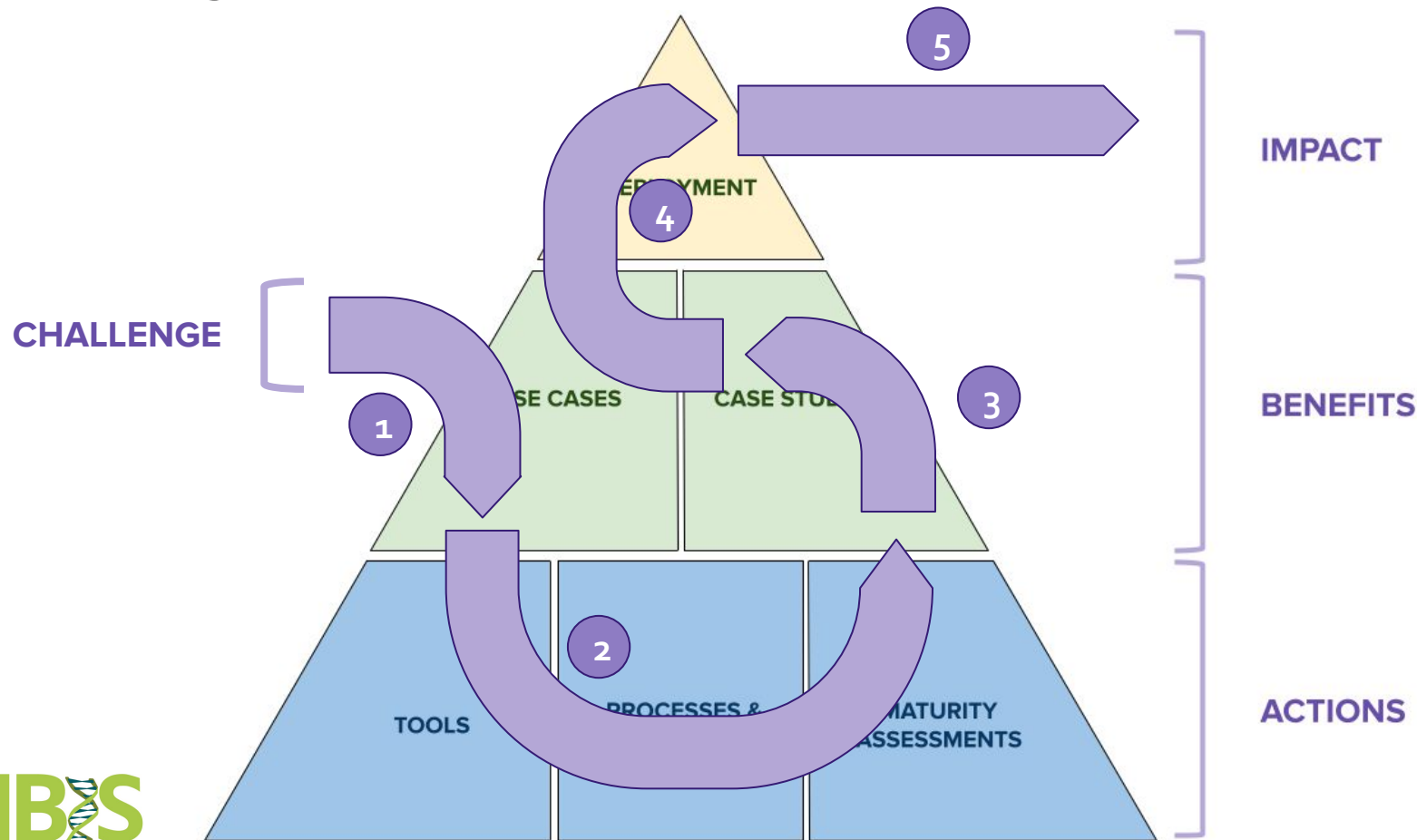


Understand community
practices, to improve
processes and create
new **products**

Telling Interoperability Stories - In Five Acts



Telling Interoperability Stories - In Five Acts



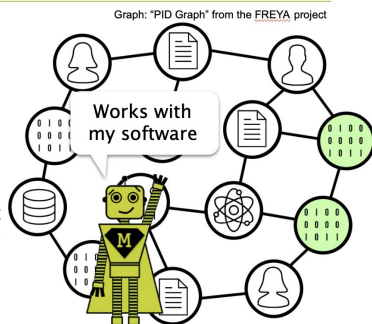
The future of data integration

Perhaps shared recipes and templates for FAIR data, tools, and services across NBIS support teams, tech groups and units?

Towards a web of FAIR data and services

"From Gutenberg to Berners-Lee"

- What types of research assets?
- Where will people be looking?
- Who should have access?
- Which standards/specifications?
- What cross-references and what documentation can you provide? (Rich metadata!)



Robot: MetaManMachine by Nikola Vasiljevic (2021),
CC BY-SA 4.0, doi:10.5281/zenodo.4471098

ELIXIR Interoperability Platform

Connecting processes to products/practices

Assemble **products** into
reusable **processes** that
drive good **practice**



Understand community
practices, to improve
processes and create
new **products**